NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

MINING DATA FROM THE ARMY RESERVE FOR ANALYSIS OF ATTRITION FACTORS

by

Robert D Radtke Jr

June 2007

Thesis Advisor: Samuel E. Buttrey
Second Reader: Roberto Szechtman

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** June 2007	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE Mining Data from the Army Reserve for Analysis of Attrition Factors		**5. FUNDING NUMBERS**	
6. AUTHOR(S) Robert D Radtke Jr			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000		**8. PERFORMING ORGANIZATION REPORT NUMBER**	
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) OCAR PAE Washington, DC		**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited		**12b. DISTRIBUTION CODE** A	

13. ABSTRACT (maximum 200 words)

The goal of this thesis was to analyze the impact of increased utilization and deployments of Troop Program Unit soldiers since 9/11, countered against the effects of demographics and of the programs and actions meant to control attrition.

This study conducted a process of data collection, data manipulation, and data-mining algorithms executed against the entire enlisted TPU population and focused toward attrition behavior.

Significant factors in determining attrition behavior included time in service, increased bonus levels and the Delayed Entry Program. Mobilizations, in and of themselves, appear to have little impact. The models we built showed significant potential for predicting behavior. We believe that this process should be continued and expanded to a tool to aid in and affect attrition.

14. SUBJECT TERMS USAR, Manpower Modeling, Enlisted Modeling, Army Reserve, Military Manpower Modeling, Data Mining, Attrition			**15. NUMBER OF PAGES** 63
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UL

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

MINING DATA FROM THE ARMY RESERVE FOR ANALYSIS OF ATTRITION FACTORS

Robert D. Radtke Jr.
Major, United States Army
B.S., United States Military Academy, 1989

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2007

Author: Robert D Radtke Jr

Approved by: Professor Samuel E. Buttrey
Thesis Advisor

Roberto Szechtman
Second Reader

James N. Eagle
Chairman, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The goal of this thesis was to analyze the impact of increased utilization and deployments of Troop Program Unit soldiers since 9/11, countered against the effects of demographics and of the programs and actions meant to control attrition.

This study conducted a process of data collection, data manipulation, and data-mining algorithms executed against the entire enlisted TPU population and focused toward attrition behavior.

Significant factors in determining attrition behavior included time in service, increased bonus levels and the Delayed Entry Program. Mobilizations, in and of themselves, appear to have little impact. The models we built showed significant potential for predicting behavior. We believe that this process should be continued and expanded to a tool to aid in and affect attrition.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I first want to thank all the Iron Majors who helped make this happen: Alison Godfrey, John Evans, John Rufenacht, Trish Ginther, and Chris Lombardi for all their assistance in collecting this data and scoping this problem and Jeff Howell for providing overwatch and a covered and concealed position.

Special thanks to Commander Kevin Maher and Colonel Andy Hernandez, whose kicks in the pants and help and guidance made sure I kept my head above water and completed this process. Also, I want to thank Lieutenant Colonel John Brau, for getting me started on this path many moons ago.

Finally, sincere thanks to Professor Sam Buttrey and Professor Roberto Szechtman for your time, work, and guidance on completing this problem and to all of my Instructors here at the Naval Postgraduate School for your time and effort in getting me to graduation! Allons!

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The goal of this thesis was to analyze the impact of increased utilization and deployments of enlisted Troop Program Unit (TPU) soldiers since 9/11, controlling for the effects of demographics and of the programs and actions meant to control attrition. Maintaining a viable and manned reserve force in this environment is critical to the security of the nation.

We conducted this study through a process of data collection, data manipulation, and data-mining algorithms executed against the entire TPU population and focused toward attrition behavior.

There were no “magic bullets” in the results. Time in service is the biggest single factor in determining attrition behavior. Increased bonus levels and the Delayed Entry Program appear to be significant factors as well. Mobilizations, in and of themselves, appear to have little impact. We hypothesize that the positive attrition trends seen within these forces is due to retention actions within the Army.

The models we built showed significant potential for predicting behavior. We believe that this process should be continued and expanded to produce a tool to aid in and affect attrition. We envision a system in which data on the service member along with responses to simple questions filled in through Army knowledge Online (AKO) or Human Resources Command could be used to focus resources and assist Retention Specialists in retaining the right people.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. BACKGROUND

The United States Army Reserve (USAR) is a key component of the Department of the Army. The Army Reserve's mission is to provide trained and ready personnel with the skills necessary to support and defend the nation during peacetime, emergencies, and war. The Selected Reserve is the foremost component to meeting this mission. The effective management of the enlisted personnel inventory in the Selected Reserve is essential for the proper support of this mission. An important component of this management is tracking and predicting attrition within the force. The Global War on Terror (GWOT) has significantly changed the utilization of Reserve Forces and has impacted the attrition behavior of the Army Reserve. A large number of actions have been taken by the Army to control this attrition.

B. THESIS OBJECTIVE

In this thesis, the goal is to determine and analyze the impact of increased utilization and deployments of TPU soldiers against the effects of various demographics and of the programs and actions meant to control attrition. The Select Reserve is that portion of the reserve that is the most ready and deployable. The Select Reserve is made up of three subsets. These subsets are the Troop Program Unit (TPU), Active Guard and Reserve (AGR), and the Individual Mobilization Augmentee (IMA). TPU soldiers are the largest portion of the Selected Reserve. They are the classic

reserve soldiers. TPU soldiers are the one week-end-a-month, two-weeks-every-summer soldiers that make up the backbone of the Army Reserve. They are the citizen soldiers; typically having civilian employment, they are the part of the force most greatly affected by deployments. They make up approximately 90% of the Selected Reserve. TPU soldiers attrite from the Army Reserve at various times and for various reasons. Some attrition is unavoidable and even positive. Therefore, we need to be able to classify attrition based on who, when, and where. For example, a TPU soldier who leaves the force after 24 months of service for unsatisfactory participation would be a “bad” attrition, whereas a soldier who leaves after 300 months due to retirement would be a “good” attrition. Also, some soldiers transfer laterally within the military, becoming active duty soldiers in the AGR program or the Regular Army, or by transferring to various other places, such as the National Guard or the Navy. These transfers, while not bad for the Nation as a whole, still place a burden on the Reserve’s accessioning agencies. Ultimately, we would like to be able to identify those factors that affect all types of attrition and use this information to predict and possibly reduce future attrition. Outcomes from this study could be used to determine recruiting and retention goals, set bonus levels, and define future manpower management programs. One possible outcome for future study is to carry this data mining process further. Corporate America uses data mining to track behaviors and predict future behavior. We have access to a much greater range of data on our audience. It is fully conceivable that this work could be carried onward to create a manpower management tool to better maintain our

force, not only to reduce attrition, but to maintain or increase levels of job satisfaction.

C. LITERATURE REVIEW

A literature review of relevant studies uncovered a CNA study entitled “Determining Patterns of Reserve Attrition since September 11, 2001. (Dolfini-Reed, 2005)” It looked at attrition across all reserve components and what their trends were since 9/11. This document was a good starting point for this analysis. It looked at attrition trends related to deployment in the Global War on Terror. It did not look at multivariate effects nor take into account any interactions that may have also been affecting attrition trends. The main factors they looked at were mobilizations, deployments, service and component, and time after deployment ended. This was a time series descriptive analysis of these factors. It did show both positive and negative trends based on these effects. The authors suggested conducting multivariate analysis and then went into what would be needed to develop a model for loss behavior. The model they suggested was a Cox Regression combined with a multinomial logit regression to create a special semi-Markov process. Ideally, this present study could be utilized in developing a multinomial factorization for support of just such a model.

We also looked at two other studies referred to by the authors of this first study. The first study, “Retention in the Reserve and Guard Components” (Hansen, 2004) looked at all reserve components from FY00 to FY03. This was a multivariate analysis, but the data set the authors used did not contain any information on mobilizations and

deployments. Some of the outcomes they found are supported by the study. They found time in service, education levels, and earning potential to be significant factors.

The second study was “Serving Away From Home: How Deployments Influence Reenlistment. (Hosek, 2002)” This study conducted an expected utility model based on a Bayesian Updating Process. They sought to model how previous deployments would affect the decision to reenlist. They used data on people facing reenlistment in the FY96-FY99 timeframe. This study was focused toward active component military and looked specifically at reenlistment, and therefore is of limited utility, but it did have some interesting conclusions. The authors of this paper found that those who had deployed were more likely to reenlist that those who did not. They hypothesized that deployments helped soldiers revise their expectations and created a bridge between past deployments and current reenlistment decisions.

II. RESEARCH METHODOLOGY

A. DATA VALIDATION

1. Data Collection and Processing

The data used in this study was provided from numerous sources. G-18, G-17, G-19 are flat files from the Total Army Personnel Database-Reserve (TAPDB-R). The G-18 contains individual personnel data on all members of the Select Reserve. The G-17 contains data on reserve units. The G-19 provides data on individual unit assignments. The AllMOB is a query file provided by USARC G-1. It contains transaction information on every individual mobilized since September 11, 2001. The Transaction File (XTX) is a file of TAPDB-R transactions that involve status changes. It is from this transaction file that we get our loss and Delayed Entry Program (DEP) data. USAR_Contract data was provided by the United States Army Recruiting Command (USAREC). It provides data on all people contracted into the Army Reserves since 1999. This includes DEP soldiers who never made it into the force. DJMSRC_EXTRACT is finance data recorded from the Defense Joint Military pay Software, Reserve Component (DJMS-RC) and extracted from the Reserve Component Management System (RCMS). It provides data on types of bonuses and years of payment. The DMDC_EXTRACT was extracted from the Defense Manpower Data Center (DMDC) Montgomery GI Bill (MGIB) database. It provides data on MGIB payments for individuals in the TPU.

The G-18 files were the most important data for this study. They provide personnel data on every individual in the TPU (Current Organization (CURORG) = H). This data set includes an entire range of demographics on personnel assigned to the Selected Reserve. These files were provided in FY chunks. We retained only TPU soldiers (CURRORG = H; MIL_PER_CA (Military Personnel Category) = E). We also removed those data elements that were sparsely populated or contained data that was not pertinent to this study (e.g. administrative data, officer specific, etc.). The FY06 file served as a starting point for creating the BASIS file. Records from FY05 with SSNs that were in the current BASIS file were removed; the remainder were then appended to BASIS. This procedure was continued through to the FY02 G-18 data. In this way, we were sure to have the most current data on any personnel in the BASIS file. After creating this file, we removed two records from this data, because they had erroneous SSNs.

Six additional tables were created from the remaining data sources. Contract_USAR was created by appending two queries provided by Recruiting Command and checked for duplicates. MOB_Count was created from the AllMOB file. The AllMOB file is a transaction file that has a separate entry for every individual mobilization of each TPU soldier. MOB_Count combines each of these based on SSN and computes the total number of days mobilized, counts number of mobilizations, indicates whether or not they were deployed, and reports information on the last mobilization. For DEPQF and LossQF, we combined five years of XTX files. We located all DEP applicants (MPAPOI (previous CURRORG) = V, MPAORG (Current CURRORG) = H) and all losses (MPAPOI = H, MPAORG <>

H); DEPQF and LossQF are the final cleaned queries with the duplicate entries removed. Where there were multiple entries, the last was used, as transactions are often amended. The loss data had six entries that were complete copies; these were removed. The DJMSRC_EXTRACT_Crosstab table was created from a crosstab query of the DJMSRC_EXTRACT data. We summed bonus amounts paid by fiscal year for each SSN. The DMDC_EXTRACT was checked for duplicates. The last transaction was used in the case of multiple entries.

A query named FinalFile pulled together these various data sets into one master table for analysis in Clementine. This query joined the six smaller tables to the BASIS file. This provided a starting database for inclusion into Clementine. Table 1 in Appendix A lays out the data dictionary for this final table.

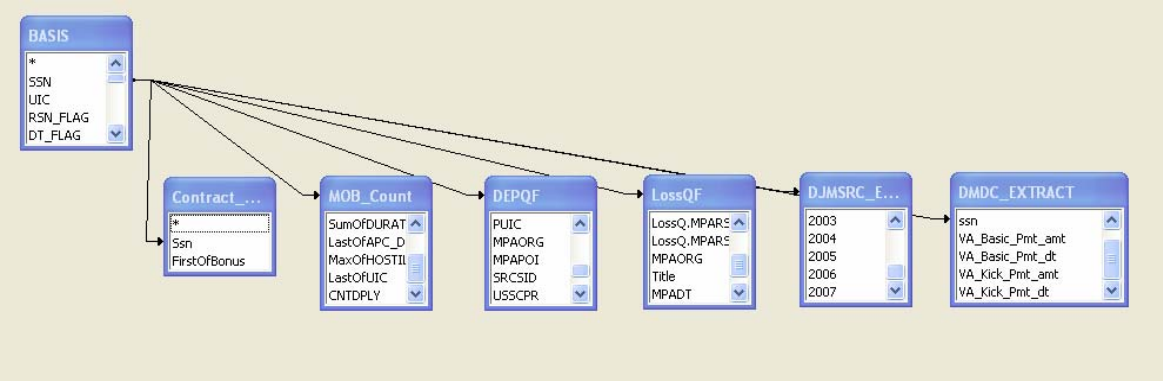


Figure 1. FinalFile Table Query View.

2. Clementine Introduction

Clementine is the SPSS enterprise-strength data mining workbench. It uses a visual interface to execute a three-

step process of reading in, manipulating, and sending data to a destination. A Clementine "Stream" is the interface that the software uses to conduct this process. A Stream consists of a various set of nodes, each of which performs a specific set of varied functions. At the simplest level, the shape of the node tells you its general function. Round nodes are source nodes. They function by grabbing data from any number of sources, including databases, Excel files, text files, or either SPSS or SAS statistical software. Hexagonal nodes are known as either field or record nodes. These nodes perform functions to prepare, transform, or otherwise modify the data in preparation for introducing it into any of the algorithms at the heart of this data-mining process. These algorithms are represented by pentagonal nodes. These nodes execute a variety of machine learning, artificial intelligence, and statistical modeling methods. They allow you to derive information from your data and create predictive models (SPSS Inc, 2006). Square nodes are output nodes. They can provide files of the transformed data for further work, as well as analytical output of the results. Triangular nodes are graphing nodes for visual analysis of the data. In order to create an acceptable model we must first continue the data manipulation process using the Clementine software.

B. DATA AGGREGATION

1. Clementine Acceptability

Data is never "clean." The previous steps were conducted to create a single data set for inclusion into Clementine. As we introduced the data to Clementine, we

began a process of accessing and modifying the data fields to ensure could be properly read by the software. This was an iterative process, stepping back and forth between Clementine, Access, and FoxPro to get the data into the right format and type. This first Clementine model (Fig. 2) was built to do just that. We used a database source node and fed it into a type node and out to a table node. We looked at outputs and ascertained a number of issues with our data. As an example, all date fields were coming across as strings. Because we were not doing a time series study, exact dates were not necessary. We fixed this by going back to Access and trimming the field to provide only the calendar year. We used a filter node to extract all personal data (SSN, Name) from our model. The first filler node put an "N/A" in all blank entries for the loss fields. The second filler was used to repair the PPSC field. This is a numeric flag field of physical profiles. The default is 111111. We also determined that using a database source link significantly slowed the execution of anything done in Clementine, so we used this stream to create a flat file for actual modeling.

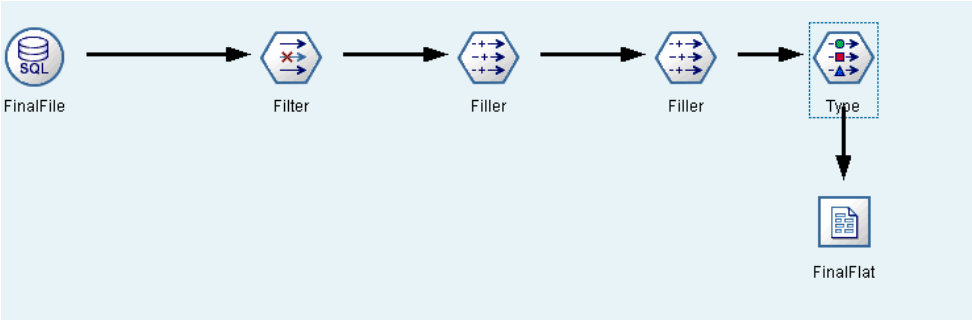


Figure 2. Initial Clementine Model.

2. Data Processing

Figure 3 shows the next set of steps. We filtered out some additional fields because they were too diversified, and would likely provide no insight. These included such things as street addresses, cities, zip codes, and grid locator codes. We then generated a new flat file with these fields removed.

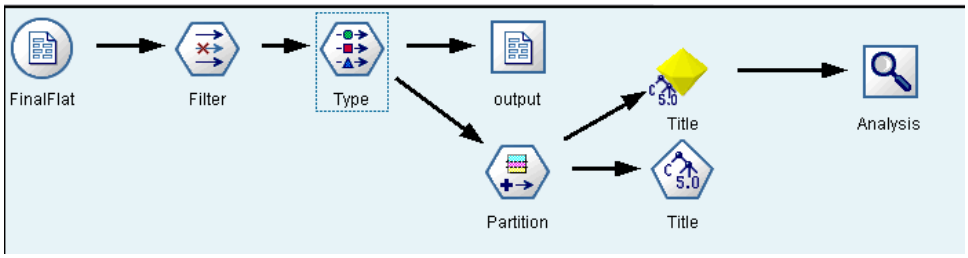


Figure 3. Initial Data Manipulation in the Model.

At this point, we began manipulating the data to prepare for modeling. The steps we followed can be seen in Figure 4. From the new output table, we added a partition set. A Partition node adds an index for the machine process. These indexes are randomly assigned to each data point, typed as either “training”, testing”, or “validation”. In the machine learning process, the algorithms will use the “training” data to develop the model. The larger the training set used, the better the model fits the existing data, but using too much training data might result in a model that over-fits the data. Because of the sheer size of the data set, we decided on a 35% training, 55% testing and 10% validation split on this data.

The next node is a Reclassification node. This took loss data from the “Title” field and created a flag field to identify losses without type.

The Binning Node was used to create a binned field based on “CIVO.” The Binning node automatically created a Derive node to implement the bins. The bin sizes were based on bins used by Army G-1 and contained in the TAPDB-R data reference guide. The Filler node was to complete this binning, as blank fields needed to be binned into “OTH.”

The next three Derive nodes provide true/false flag indicators to create the fields “Profile”, “Mobilized” and “Deployed.” The next Derive node provides the field “TISatLOSS,” which is the number of years in service that any losing service member had when they left. “RRC” is the two digit lead for the major command of the individual. “CMF” is the first two digits of the service member’s primary MOS.

The first Filler node places an H in all blank DEPPER fields. Any blanks indicated an MPAOI of H, rather than V, meaning they entered through the Delayed Training Program (DTP), instead of the Delayed Entry Program (DEP). The second filler placed an M (default) in blank entries of the HT_WT_IND data field.

One final Reclassification node, entitled Destination, created a more generalized output field for analysis, based on some initial outcomes from test runs. This final reclassification allowed me to type individuals based on “TPU” (Those who have stayed), “OUT” (Individuals who left to a less ready status, such as the IRR or had a complete break from service), “MILITARY” (Individuals who went to

another form of service; this includes IMA, AGR, RA, Service Academy, or other branch of service), and “RETIRED” (those who completed full terms of service and were retired in one form or another).

As a last node for data manipulation, we added a Select node. This node was set to look at soldiers who enlisted since 9/11. We set this as an adjunct to run models on that special part of the population.

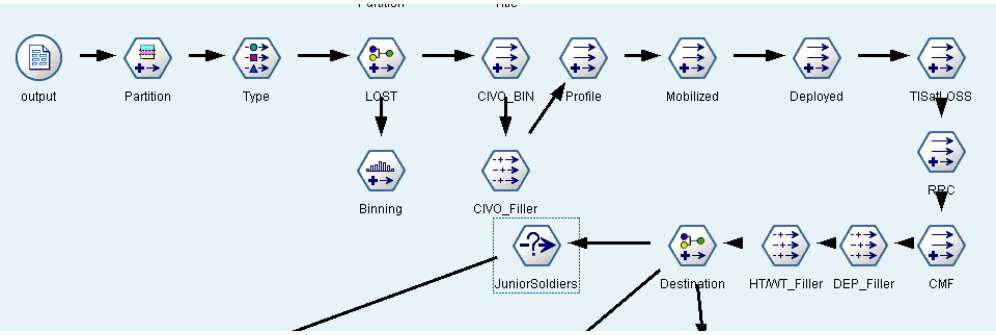


Figure 4. Data Manipulation in the Clementine Model.

3. Data Classification

Type nodes are used to set the role of the data for use by the models. These nodes are used to specify field metadata and properties. These data include type, label, direction, and values. They are the final control to set the input of the data into the model (SPSS Inc., 2006).

The Tables in appendix A show the final settings of the Type nodes. Grey fields were excluded from the model. Some of these fields are excluded for reasons discussed earlier. Many others were excluded because they were too aggregated

and provided no real information. All the other output fields were considered and rejected during the modeling process. Finally, some were excluded because they had dependent values that allowed the model to "cheat." One example was the RSC field. Because RSCs changed to RRCs during the early part of this time window, any record with a "RSC" value in the RSC field was an obvious loss, corresponding to a soldier who had not been in a unit for at least 4 years. Fields marked with an asterisk (*) were possible output fields. The yellow field (Destination) was the final output field.

C. MODELING METHODOLOGY

We began modeling the data using association rule models. Association rules are statements in the form of "if *antecedents* then *consequences.*" We chose this model to find hidden or unanticipated associations in the data, such as fields that were actually output-dependent. We used the Apriori node to look at association rules in the data. This algorithm was run at various points in the model while using either "Title," "Loss," or "Destination" as the output. The settings were at 5% support with 80% confidence. Outputs from the algorithm were used to eliminate some of the fields in the data set. The RSC field was one such example.

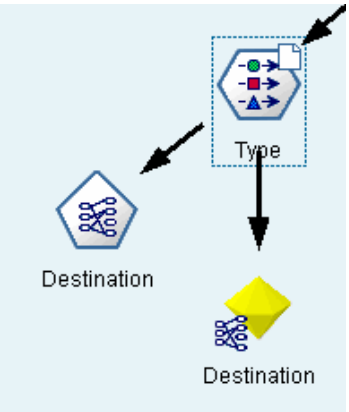


Figure 5. Apriori Node in Clementine Model.

For this study, we focused on decision tree models. A decision tree allows us to see how different factors and interactions occur in relation to the outcomes we are trying to determine and analyze. There are four types of decision tree algorithms in Clementine. They are Classification and Regression (C&R), CHAID, QUEST, and C5.0. CHAID and QUEST produced the weakest results and were rejected.

C&R and C5.0 both partition data recursively along splits generated from the outcomes in the training data. The C&R tree finds an "impurity index" in the data and looks for the split that provides the greatest reduction in that index. The C&R tree will only provide binary splits. The C5.0 tree works in the same fashion, but adds a couple of features. C5.0 is not limited to a binary split, and can utilize "boosting." "Boosting" allows the algorithm to build multiple successive models. Each successive model attempts to repair the errors in the previous model, and then allows all of these models to "vote," providing a boosted prediction (SPSS Inc, 2006). The advantage of the C5.0 algorithm is better demonstrated prediction at the cost

of a much more complicated tree. The C&R tree gives a much easier to comprehend outcome but with a lower success in prediction. Figure six displays the model set-up for both the Full set and the Junior Enlisted set.

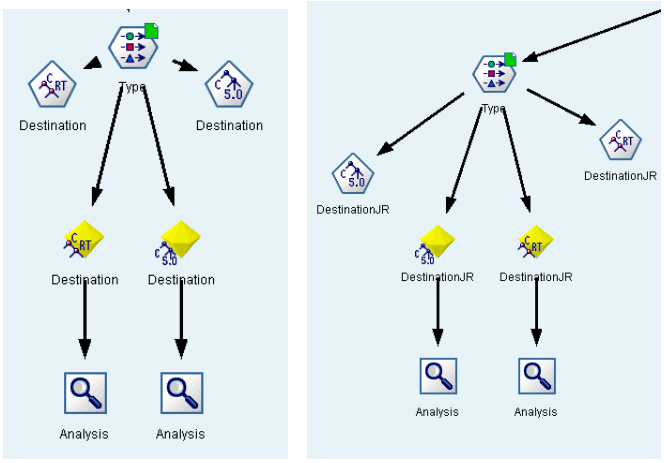


Figure 6. Model Execution.

THIS PAGE INTENTIONALLY LEFT BLANK

III. DISCUSSION AND FINDINGS

A. PARTIAL RUNS

As discussed in the previous chapters, we conducted some interim modeling as we went through the development of the final model structure. We used the Apriori algorithm and the C&R tree to look at results along the way. These algorithms ran in a relatively rapid manner (minutes rather than hours), and we were able to use them to find input data fields that were dependent on response fields. These dependent fields needed to be excluded from the model.

We conducted experimental runs with Neural Nets, which produced significantly lower prediction success. This may have been a function of not allowing the runs to complete, but after 19 hours, the most successful model was still only at approximately 58% prediction accuracy.

B. C&R TREE OUTCOMES

1. Full Data Set

The following graph is the final decision tree produced for the entire data set, using the C&R algorithm. As an example, the algorithm found the biggest split reduction at a total bonus amount of $2985. Going down the right side of the tree, those with a bonus greater than $2985 are checked for DMOSQ. Those with a DMOSQ of A or X are predicted to get out. This makes sense. A person with an A or X has usually failed to complete training and will most likely not receive his or her bonus. Of those with the other DMOSQ

codes, the model then makes cuts based on Family Care Plan status, time in the reserves, and deployment status. Notice that many splits do not change the decision of the model, but do modify the confidence of the results.

Figure 7. C&R Decision Tree (Entire Population)

This model provides the following results

'Partition'	1_Training		2_Testing		3_Validation	
Correct	74,545	70.99%	148,420	70.95%	24,597	70.81%
Wrong	30,470	29.01%	60,756	29.05%	10,142	29.19%
Total	105,015		209,176		34,739	

Table 1. C&R Results

The 71% accuracy should be compared to the naïve model, which simply predicts the most common outcome for every observation. In this case, the naïve model would predict TPU for all cases and be correct 52% of the time. Notice that this model never makes a prediction of Military, but appears to be splitting that category between the Out and TPU categories. Most critical, those who actually remained in the TPU were correctly predicted to do so about 88%.

'Partition' = 1_Training	Out	Retired	TPU
Military	7,499	152	5,569
Out	22,918	767	8,781
Retired	360	3,438	1,134
TPU	4,868	1,340	48,189
'Partition' = 2_Testing	Out	Retired	TPU
Military	14,796	319	11,044
Out	46,167	1,459	17,769
Retired	768	6,580	2,166
TPU	9,872	2,563	95,673
'Partition' = 3_Validation	Out	Retired	TPU
Military	2,437	60	1,843
Out	7,543	247	2,884
Retired	124	1,057	351
TPU	1,752	444	15,997

Table 2. C&R Predictions vs. Outcomes

2. Junior Soldier Data Set

The first model generated for the junior soldier set (Figure 8) was accurate but uninformative. It tells us that not completing training is the factor most correlated with loss in this junior enlisted set. The Army Reserve has already recognized this issue. This was one of the main reasons for implementing the DEP. A large percentage of the soldiers entering into the Army Reserve never complete training and are never a viable asset to the Army. The DEP was meant to prevent these undeployable assets from being accessed into the Force until they leave for training.

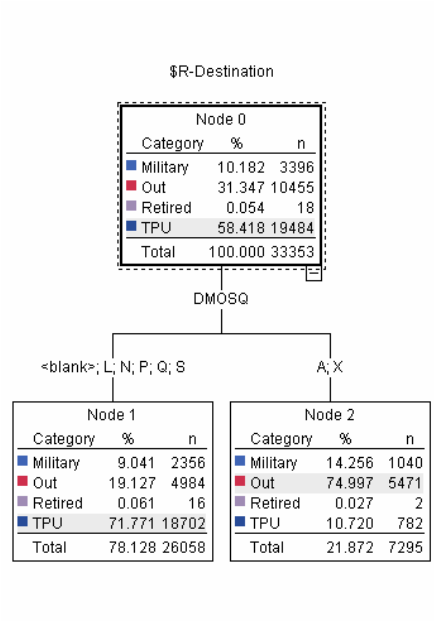


Figure 8. C&R Decision Tree (Junior Soldiers)

We reran the model by excluding this factor from this data set. This produced the following tree (figure 9). This provided further insight into other factors affecting attrition. The prediction results for this tree were slightly less accurate than those of the previous tree.

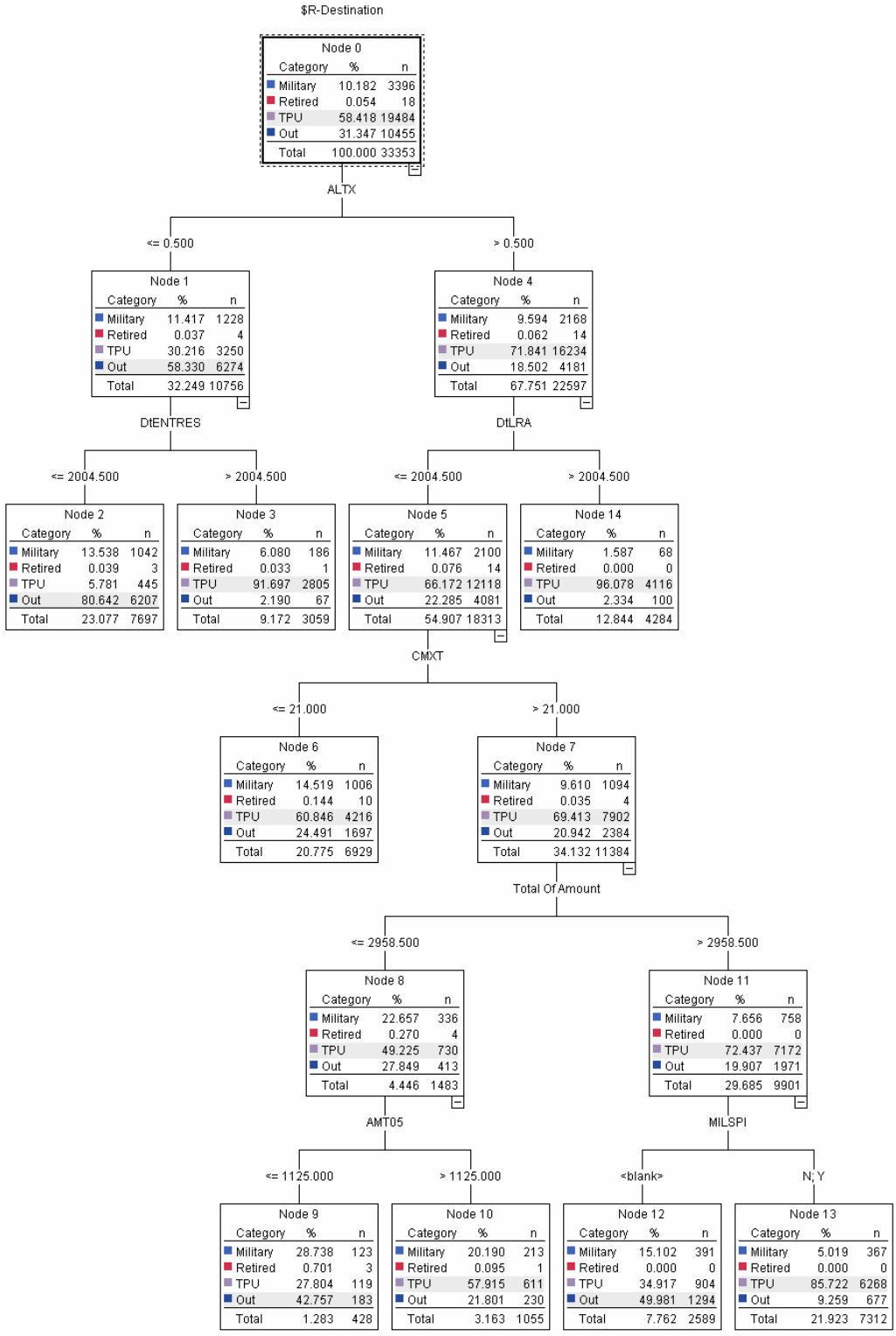


Figure 9. Second C&R Decision Tree (Junior Soldiers)

The results for this tree are as follows (Table 3). The predicted accuracy of this tree was less than one percent lower than that of the first tree. The 72% accuracy found here should be compared to the 58% accuracy of the naïve model.

'Partition'	1_Training		2_Testing		3_Validation	
Correct	24,173	72.48%	48,191	71.92%	8,096	72.47%
Wrong	9,180	27.52%	18,812	28.08%	3,076	27.53%
Total	33,353		67,003		11,172	

Table 3. Junior Soldier C&R Results

This model, looking at just junior soldiers, would (and should) never predict a retirement. The retirements that are in this data set are medical in nature. Looking at Table 4, we see that this model does a very good job of predicting those who stayed in the TPU, but noticeably worse in predicting who would get out. The real gain in this model, though, is that underlying this is data on how strong the prediction is. The decision tree shows percentages at the end nodes. These can be interpreted as predicted conditional probabilities to build specific prediction models on how many people we are at risk of being lost, based on the composition of the force. In this way, we are able to classify individuals by their level of risk. Table 4 breaks out the predictions for this model against actual performance for the three partitioned sets.

'Partition' = 1_Training	Out	TPU
Military	1,040	2,356
Out	5,471	4,984
Retired	2	16
TPU	782	18,702

'Partition' = 2_Testing	Out	TPU
Military	2,006	4,838
Out	11,035	10,298
Retired	5	20
TPU	1,645	37,156

'Partition' = 3_Validation	Out	TPU
Military	329	816
Out	1,798	1,641
Retired	2	6
TPU	282	6,298

Table 4. Jr. Soldier C&R Predictions vs. Outcomes

C. C5.0 TREE OUTCOMES

The trees generated from the C5.0 algorithm are much larger and much more complicated than those of the C&R algorithm. Conversely, they were also significantly more informative and had better prediction outcomes. Because of their complexity, we are unable to display the decision trees in this document.

1. Full Data Set

The full set model used 8 trees to boost its results and provided up to 14 levels of significant factors in each tree. The most prevalent factors correlated with attrition included time in service, unit, the delayed entry program, dependency, and education. Many other factors appeared but were much less prevalent. The prevalence of a factor was

determined by the number of times it appeared and how high in the tree it appeared. The results for this model are listed below.

'Partition'	1_Training		2_Testing		3_Validation	
Correct	83,982	79.97%	164,465	78.63%	27,348	78.72%
Wrong	21,033	20.03%	44,711	21.37%	7,391	21.28%
Total	105,015		209,176		34,739	

Table 5. C5.0 Results

The data in these tables shows the improved performance of the C5.0 over the C&R tree for modeling this data.

'Partition' = 1_Training	Military	Out	Retired	TPU
Military	2,867	7,378	223	2,752
Out	1,048	26,100	836	4,482
Retired	3	295	4,006	628
TPU	325	2,711	352	51,009

'Partition' = 2_Testing	Military	Out	Retired	TPU
Military	4,956	14,934	469	5,800
Out	2,447	51,710	1,670	9,568
Retired	14	678	7,407	1,415
TPU	798	6,153	765	100,392

'Partition' = 3_Validation	Military	Out	Retired	TPU
Military	849	2,462	80	949
Out	417	8,401	285	1,571
Retired	3	107	1,189	233
TPU	155	1,005	124	16,909

Table 6. C5.0 Prediction vs. Outcomes

2. Junior Soldier Data Set

The Junior Enlisted set model used 10 trees to boost its results and provided up to 16 levels of significant factors in each tree. The most prevalent factors included

unit, delayed entry program, DMOSQ, and marital status. Many other factors appeared but were much less prevalent. The results for this model are listed below.

Results for output field Destination

'Partition'	1_Training		2_Testing		3_Validation	
Correct	29,657	88.92%	57,049	85.14%	9,537	85.37%
Wrong	3,696	11.08%	9,954	14.86%	1,635	14.63%
Total	33,353		67,003		11,172	

Table 7. Jr. Soldier C5.0 Results

'Partition' = 1_Training	Military	Out	Retired	TPU
Military	965	1,886	0	545
Out	117	9,777	0	561
Retired	1	6	2	9
TPU	63	508	0	18,913

'Partition' = 2_Testing	Military	Out	Retired	TPU
Military	950	4,403	0	1,491
Out	729	18,948	0	1,656
Retired	4	13	0	8
TPU	204	1,445	1	37,151

'Partition' = 3_Validation	Military	Out	Retired	TPU
Military	170	730	0	245
Out	131	3,050	0	258
Retired	0	6	0	2
TPU	22	241	0	6,317

Table 8. Jr. Soldier C5.0 Prediction vs. Outcomes

THIS PAGE INTENTIONALLY LEFT BLANK

IV. CONCLUSIONS AND RECOMMENDATIONS

This study can provide the framework to support the type of Markov process prediction model that the authors of the "Patterns of Reserve Attrition" study (Dolfini-Reed, 2005) talked about. Additionally, because it provides information about individuals, this process can be used for developing a retention tool, in order to help Retention NCOs and Commanders better identify who may be at risk and focus limited resources toward maintaining the force.

There is not as much insight into factors as one would have hoped. Time in service was, unsurprisingly, the largest factor in determining attrition behavior. Every model we generated using the full data set found that behavior seemed to change at 12 to 14 and again at 18 years of service. The C5.0 model found a distinct and positive split around the Delayed Entry Program. Bonus amounts at $2985 and $3500 were also significant in both sets of data. MGIB data was an almost non-existent factor in any of the trees. On reflection, we believe this may actually be a function of not having the right data for this field. What may be better data is who is enrolled in the program, rather than who is receiving benefits. During our time with Recruiting Command, it was a commonly held belief that the MGIB was the most cost-efficient of any of the available benefits, as a large number of people signed up because of it, but a much smaller number of people actually ever used it. Surprisingly, the mobilization, combat, or other "go to war" indicators seemed to be insignificant factors in attrition. When they did show up in a model, they displayed

a positive trend with increased usage. This supports the trends shown in the other studies considered. Therefore, it may be tempting to infer that the war is not having a negative impact on the decision of individuals to remain serving in the Army Reserve. The generally negative outcomes from post-mobilization surveys and the downward trend in propensity to join the military that the Army seems to be experiencing (personal communications, 2006) would seem to indicate, however, that the potential for attrition should increase as well. The data from this study provides some support to the conclusions of Hosek and Totten (2002), which hypothesized that deployments helped to vest interest in service and reduced naïveté in military service. Although this data neither supports nor refutes that claim, our theory is that the Army is currently and has in the past, proactively attacked what was seen as a potential manpower problem and effectively eliminated it.

We believe there is significant potential for expansion and follow on work from this thesis. Conducting similar studies against the remainder of the Select Reserve should be equally informative. Further regression models might predict time in service when a loss occurs. Another critical step in developing the proposed model in the Dolfini-Reed paper would be to conduct time series analysis with these data for support the Markov process model.

The models we did build showed significant potential for predicting behavior. We believe that this process should be continued and expanded to a tool to aid in and affect attrition. We envision a system in which data on the service member along with responses to simple questions

filled in through Army Knowledge Online (AKO) or Human Resources Command could be used to focus resources and assist Retention Specialists in retaining the right people. Additionally, we encountered difficulty with obtaining data. In some cases, parochialism, proprietary attitudes, and “stovepiping” prevent data from being available to the analytic cells we worked with in the Army Reserve. Some of the data manipulations in this study had never been done before, because of just these problems. There are many other data sets out there that could possibly improve on this study as well. Some data sets we know to exist but were unable to use include tuition assistance and retirement points.

A final recommendation would be to set up a data warehouse for the Army Reserve analytical cells. This might perhaps be controlled by OCAR-PAE or USARC, to act as single source of study data for the Army Reserve. One possible solution would be to integrate these other data into TAPDB-R or RCMS.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A DATA DICTIONARY

Table 9. Data Dictionary

DBF NAME	DATA REFERENCE NAME	DESCRIPTION	OWNING USARC/USAR AGENCY
SSN	SOCIAL SECURITY NUMBER	A soldier's Social Security Number	DCSPER
UIC	CURRENT UIC	The Unit Identification Code that identifies the organizational assignment for a soldier	DIRFP
RSN_FLAG	SUSP FAVOR PERS ACTION REASON	The reason for suspending favorable personnel actions (flag) for a soldier	DCSPER
DtF	DATE SUSP FAVOR PERS ACTION	Year of suspension of favorable personnel actions (flag) FORMAT: CCYY	DCSPER
NAME	NAME INDIVIDUAL	A soldier's full name	DCSPER
MIL_PER_CA	MILITARY PERSONNEL CLASS	The categories into which the soldiers of the Armed Forces are divided based upon their grade and status (Commissioned Officer (CO), Warrant Officer (WO), Enlisted (ENLD), Academy Cadet) (Limited to E only)	DCSPER
DtBiY	DATE OF BIRTH	The year a soldier was born FORMAT: CCYY	DCSPER
HIGH	HEIGHT INDIVIDUAL	A soldier's actual height, in inches, as indicated during the most current weigh-in or medical examination	SURGEON
WGNT	WEIGHT INDIVIDUAL	A soldier's actual weight, in pounds, as indicated during the most current weigh-in or medical examination	SURGEON
HT_WT_IND	HGT/WT ACCEPTABILITY INDIC	Indicates whether or not a soldier's weight is acceptable for the soldier's height, even if not within the Army prescribed limits	SURGEON
SEX	SEX	The sex of a soldier	SURGEON
ETH_GP	ETHNIC GROUP	A soldier's ethnic group (A segment of the population that possesses common characteristics and a cultural heritage significantly different from that of the general U.S. population and closely identifies with that cultural heritage)	DCSPER
RACE	RACE/POPULATION GROUP	A soldier's race (A division of the human population having descent or origin in particular peoples or racial groups)	DCSPER
CITZ	CITIZENSHIP STATUS	The legal (statutory) origin of a soldier's United States citizenship status	DCSPER
MAST	MARITAL STATUS	A soldier's legal marital status	DCSPER
DEPN	NUMBER OF DEPENDENTS	The number of dependents for a soldier (Dependents: Persons for whom the sponsor (normally the head of a household) provides support in accordance with the provisions of DOD Military Pay and Allowance Entitlements manual)	DCSPER
RELI	RELIGIOUS DENOMINATION	A soldier's religious denomination (A sect or group of individuals with similar theological beliefs)	CHAPLAIN
CIVO	CIVILIAN OCCUPATIONAL CATEGORY	The general category into which a soldier's civilian occupation is classified based on the type of work performed	DCSPER
STREET	STREET ADDRESS	The street address portion of a soldier's address	CIO
CITY	ADDRESS CITY	The name of the city in a soldier's address	CIO
STATE	STATES/TERRITORIES	The name of the state in a soldier's address	CIO

DBF NAME	DATA REFERENCE NAME	DESCRIPTION	OWNING USARC/USAR AGENCY
	OF US		
ZIP	ZIP CODE	A 5-/9-digit zip code in a soldier's address	CIO
ZGLC	GRID LOCATOR CODE	A code denoting a specific geographic location within the boundaries of the Continental US; developed by using WIAP-Z to divide the Continental US into quadrants 15 miles square (000000 - 995995)	SYS GEN
GRAS	GRADE ARMED SERVICE	A soldier's grade (A rating in a graduated progression of ratings in an Armed Service; this rating is equal to a grade level or is in a relative position between grade levels within the United States hierarchy of grades)	SYS GEN
GRADE	GRADE TITLE - US ARMY	The 3-character abbreviation of the rank a soldier holds in the United States Army (COL, CPT, CW3, SGT, PV1)	DCSPER
DtR	DATE OF RANK - RESERVE	The year a soldier's rank in the reserves became effective - This date establishes the relative seniority of a soldier among others who possess the same Reserve military grade FORMAT: CCYY	DCSPER
DtPEBD	PAY ENTRY BASIC DATE	The date that establishes the beginning of a soldier's creditable service for pay purposes (Equals the date of enlistment for NPS gains) FORMAT: CCYY	DCSCOMPT
DtEXP	EXPIRATION STATUTORY MIL OBLG DATE	The date when a soldier has completed or will complete a period of service required by statute (The initial period of service, active or reserve, required by statute is 8 years) FORMAT: CCYY	OCAR RTD
DtTPUEXP	EXPIRATION OF TPU SERVICE DATE	The date indicating the expiration of the period a soldier is currently obligated or expected to serve as a member of the Selected Reserve with either a Reserve unit (TPU) or on an active duty tour (AGR) FORMAT: CCYY	OCAR RTD
DtLRA	DATE LAST RELEASED ACTIVE DUTY	The date a soldier last completed a period of active duty or active duty for training (Non-TPU training AD) FORMAT: CCYY	DCSPER
ACT_FEDSVC	NUMBER MONTHS ACT FED SVC	A soldier's cumulative, creditable period of full-time active duty, expressed in 30-day increments (Includes periods of AT, ADT, ADSW, IADT, etc.)	DCSPER
PPSC	PHYSICAL PROFILE SERIAL	(PULHES) - An estimate of the overall ability of a soldier to perform military duties by consideration of the physical and mental condition (PULHES consists of six numbers - each from 1-4 - indicating a rating for the soldier in each of the following categories: Physical Capacity Indicator (P), Upper Extremities Capacity Indicator (U), Lower Extremities Capacity Indicator (L), Hearing/Ears Capacity Indicator (H), Eyes/Vision Capacity Indicator (E), Psychiatric Capacity Indicator (S) in that sequence - (Example: 111111 indicates no limitations in any category)	SURGEON
PHCC	PHYSICAL CATEGORY	Represents certain combinations of physical profile serial codes (PULHES) and the most significant duty limitations	SURGEON
APFT_IND	APRT INDICATOR	Designates that a soldier passed or failed the last performance of the Army Physical Readiness Test	DCSOPS
DEPL	PERS DEPLOYABILITY LIMITATION	The most significant factor which precludes the overseas assignment of a soldier during full mobilization	DCSPER
MILED_COMP	MIL EDUC	The completion status of a soldier's military professional	DCSOPS

DBF NAME	DATA REFERENCE NAME	DESCRIPTION	OWNING USARC/USAR AGENCY
	COMPLETED STATUS	training	
CIED	CIVILIAN EDUC CERT COMPLETED	The highest level of formal academic education, in approved program of study at a non-military institution or service academy, attained by a soldier (Completion should be recognized or certified by a diploma, degree, document, or other certificate)	DCSOPS
CVEL	CIVILIAN EDUCATION LEVEL	The highest level of formal academic (non-military) education obtained by a soldier	DCSOPS
MSCE	MAJOR SUBJECT COLLEGE EDUC	A soldier's major field of study for the highest civilian education attained	DCSPER
DtENTRY	YR-MO INITIAL ENTRY MIL SVC	The year and month a soldier was first commissioned or enlisted in any military service of the United States (Active or Reserve) - This date is fixed and is not adjusted for breaks in service FORMAT: CCYY	DCSPER
DtENTRES	YR-MO INITIAL ENTRY RES	The year and month a soldier affiliates or enlists in any Reserve component (non-EAD) for the first time - This year and month is fixed and would not be adjusted for breaks in Reserve service (For non-prior service members, this year and month would equal the year and month of initial entry military service - often blank for pre-reservist if not entered from OMPF) FORMAT: CCYYMM	DCSPER
AFSG	AFQT SCORE GROUPS	The aggregated percentile test score group into which a soldier's score on the Armed Forces Qualification Test falls	DCSPER
AFQT	AFQT PERCENTILE SCORE	The percentile score attained by an examinee on the Amed Forces Qualification Test	DCSPER
DtETS	EXPN READY RESERVE OBLG DATE	A date indicating the expiration of the period an enlisted soldier is required by law or contractual agreement to serve as a member of the Ready Reserve (TPU, AGR, Control Group) FORMAT: CCYY	OCAR RTD
NEXE	NBR OF ENLISTMENT EXTENSIONS	The number of extensions associated with a soldier's current enlistment	OCAR RTD
CMXT	CUMULATIVE MONTHS EXTENSION	The total (cumulative) number of months a soldier has extended his/her current ready reserve obligation	OCAR RTD
PMOS	PRIMARY MOSD - ENLISTED	The Military Occupational Specialty Designator (MOSD) of an enlisted soldier that is of first significance to the Army in terms of training, experience, demonstrated qualifications, and Army needs	DCSPER
SKLVL	SKILL LEVEL	Level of proficiency required for performance of a specific military job, and the level of proficiency at which a soldier qualifies in the Military Occupational Specialty (MOS) (The 4th character in a Primary MOSD - Enlisted)	DCSPER
SMOS	SECONDARY MOSD - ENLISTED	Identifies a Military Occupational Specialty Designator (MOSD) of an enlisted soldier that is next in significance to the primary MOSD - Enlisted	DCSPER
AMOS	ADDITIONAL MOSD - ENLISTED	Designates a Military Occupational Specialty Designator (MOSD) that is in addition to the primary and secondary MOSDs	DCSPER

DBF NAME	DATA REFERENCE NAME	DESCRIPTION	OWNING USARC/USAR AGENCY
ASI	ASI - COMMISSIONED OFFICER ASI - ENLISTED ASI - WARRANT OFFICER	Commissioned Officer (CO) - An Additional Skill Identifier (ASI) indicating a specialized skill that is required to perform the duties of a position but is not necessarily related to any one particular specialty Enlisted (ENLD) - An Additional Skill Identifier (ASI) indicating a specialized skill closely related to or an adjunct to that required by an enlisted MOS Warrant Officer (WO) - An Additional Skill Identifier (ASI) indicating a specialized skill or equipment unique to a position to identify those qualified for a position	DCSPER
GOOD_YRSVC	TOTAL SATISFACTORY YEARS RET	The number of years of military service that a soldier is credited with having served that are acceptable for retirement purposes	DCSPER
DSSI	DUTY POSD	Specifies the duty that a soldier is actually performing (Consists of the soldier's MOS, a First Duty ASI, and either a Second Duty ASI or a Duty Language Identifier	DIRFP
UCAG	USAR COMMAND OF ASSIGNMENT	An organization in the United States Army Reserve that is normally commanded by a General Officer and responsible for units within its command structure or within a specified geographical boundary	DIRFP
FCPSCD	INDIVIDUAL FAMILY CARE PLAN	Indicates the status of the arrangements required of sole parents or military couples to provide for their dependents while involved in wartime duties	DCSPER
FCPSDT	FAMILY CARE PLAN SUBMISSION DATE	The most recent date a Family Care Plan was submitted	DCSPER
SOPTDD	SOLE PARENT DEPENDENT DESIGN	Designates a soldier as the sole parent of a dependent	DCSPER
MILSPI	MILITARY SPOUSE INDICATOR	Indicates that a soldier's spouse is also in the military	DCSPER
MUSARC	MAJOR USAR COMMAND ASG	A Reserve Command directly subordinate to, and constituting a major mission element of, a major Army subcommand (A numeric 1st position indicates the US Army)	DIRFP
DtADTE	ACCESSION DATE	Actual date a soldier was gained into the current reserve component category FORMAT: CCYY	DCSPER
ORIG	ORIGINATOR CODE	A code to uniquely identify each originator submitting data to the system (Consists of the Data Entry Point (MUSARC code) + the Originator Designator (specific office w/in an agency) + the Data Entry Clerk (specific user id))	DCSPER
DMOSQ	DUTY QUALIFICATION CODE	A code indicating the Commander's evaluation of the ability of the soldier's qualification to perform the duties of the assigned position as defined by AR 140-185, Table 1-1	DCSOPS
UNITNAME	UNIT NAME	The name of the unit to which a soldier is assigned	DIRFP
TIER		Unit Priority	DCSOPS
MSCNAME		MSC Name	DIRFP
RSC		RSC Name	DIRFP
PRI	FIRST 3 CHAR OF SM PRIMARY SPEC	G18CWE.DBF ONLY - FIRST 3 CHAR OF SM PRIMARY SPEC	DCSPER
PRIX	FOURTH CHAR OF SM	G18CWE.DBF ONLY - FOURTH CHAR OF SM PRIMARY	DCSPER

DBF NAME	DATA REFERENCE NAME	DESCRIPTION	OWNING USARC/USAR AGENCY
	PRIMARY SPEC	SPEC	
SEC	FIRST 3 CHAR OF SM SECOND SPEC	G18CWE.DBF ONLY - FIRST 3 CHAR OF SM SECOND SPEC	DCSPER
SECX	FOURTH CHAR OF SM SECOND SPEC	G18CWE.DBF ONLY - FOURTH CHAR OF SM SECOND SPEC	DCSPER
ALT	FIRST 3 CHAR OF SM ALT SPEC	G18CWE.DBF ONLY - FIRST 3 CHAR OF SM ALT SPEC	DCSPER
ALTX	FOURTH CHAR OF SM ALT SPEC	G18CWE.DBF ONLY - FOURTH CHAR OF SM ALT SPEC	DCSPER
ASVABCL	ASVAB - CLERICAL	The score earned by a soldier on the Clerical portion of the ASVAB	DCSPER
ASVABCO	ASVAB - COMBAT ORIENTATION	The score earned by a soldier on the Combat Orientation portion of the ASVAB	DCSPER
ASVABEL	ASVAB - ELECTRICAL	The score earned by a soldier on the Electrical portion of the ASVAB	DCSPER
ASVABFA	ASVAB - FIELD ARTILLERY	The score earned by a soldier on the Field Artillery portion of the ASVAB	DCSPER
ASVABOF	ASVAB - FOOD SERVICE	The score earned by a soldier on the Food Service portion of the ASVAB	DCSPER
ASVABGT	ASVAB - GENERAL TECHNICAL	The score earned by a soldier on the General Technical portion of the ASVAB	DCSPER
ASVABGM	ASVAB - GENERAL MAINTENANCE	The score earned by a soldier on the General Maintenance portion of the ASVAB	DCSPER
ASVABMM	ASVAB - MOTOR MAINTENANCE	The score earned by a soldier on the Motor Maintenance portion of the ASVAB	DCSPER
ASVABSC	ASVAB - SKILL COMMUNICATIONS	The score earned by a soldier on the Skill Communications portion of the ASVAB	DCSPER
ASVABST	ASVAB - SKILL TECHNICAL	The score earned by a soldier on the Skill Technical portion of the ASVAB	DCSPER
DtEFFDG	EFFECTIVE DATE OF GRADE	The date a soldier's grade became effective: CCYY	DCSPER
FirstOfBonus	N/A	Bonus Info from Recruiting Command	USAREC
*LossTyp	MPA Type from XTX	Code describing Type of loss	XTX
*LossTypDesc		Description of MPA Type	XTX
LossRsn	MPA Reason from XTX	Code describing Reason of loss	XTX
*LossRsnDesc		Description of MPA Reason	XTX
*MPAORG	CurrOrg for Loss	Code describing Loss Destination	XTX
*Title		Description of Loss Destination Code	XTX
*DtLOSS	Year Loss Occurred		XTX
DEPPER	CurrOrg orig	IDs personnel who were in DEP	XTX
LastOfDMOS	DMOS for mobilized pers.	DMOS for last Mobilization	ALLMOB
SumOfDURATION		# days mobilized	ALLMOB
LastOfAPC_DESC		Operation last mobilized for	ALLMOB
CombatFLG		Count of mobilizations with Hostile Fire Pay	ALLMOB
LastOfUIC		Last UIC mobilized for	ALLMOB

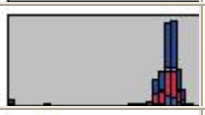
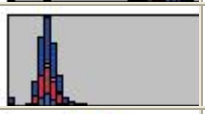
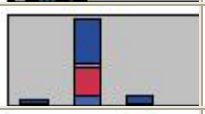
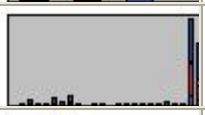
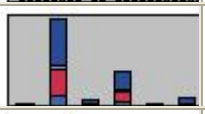
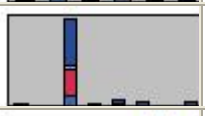
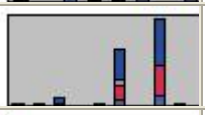
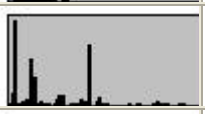
DBF NAME	DATA REFERENCE NAME	DESCRIPTION	OWNING USARC/USAR AGENCY
CNTDPLY		# of times Mobilized	ALLMOB
Total Of Amount		Total Bonus Amount since FY 1996	DJMSRC
AMT02		Bonus Amount for that FY	DJMSRC
AMT03		Bonus Amount for that FY	DJMSRC
AMT04		Bonus Amount for that FY	DJMSRC
AMT05		Bonus Amount for that FY	DJMSRC
AMT06		Bonus Amount for that FY	DJMSRC
BasMGIB		Amount paid from VA for ed benefits	DMDC
DtMGIB		Last FY of Ed Benefits	DMDC
KicMGIB		Amount paid from VA for ed benefits (kicker)	DMDC

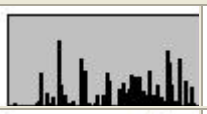
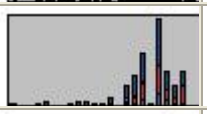
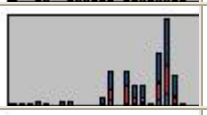
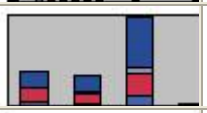
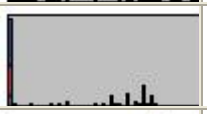
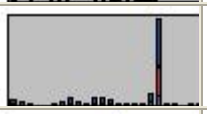
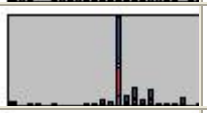
Table 10. Data Dictionary (Calculated Fields)

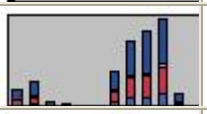
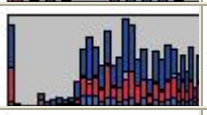
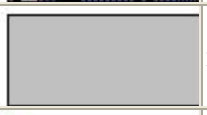
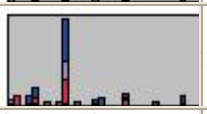
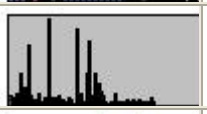
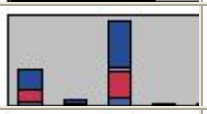
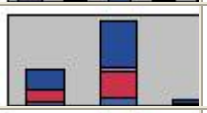
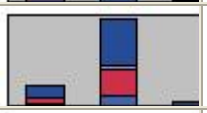
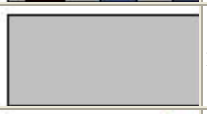
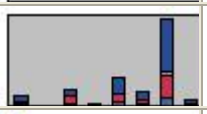
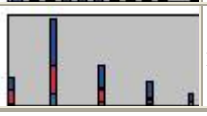
DBF NAME	DESCRIPTION	OWNING
Partition	Random generated field to separate training, testing and validating data	CALCULATED
*LOST	Generated from XTX files (flag for loss)	CALCULATED
CIVO_BIN	BIN of civilian occupations per TAPDB-R Descriptions	CALCULATED
Profile	Flag to indicate a permanent profile	CALCULATED
Mobilized	Flag indicating mobilization since 9/11	CALCULATED
Deployed	Flag indicating deployment to a warzone since 9/11	CALCULATED
*TISatLOSS	# of Years of service at time a loss occurred	CALCULATED
RRC	2 digit indicator of RRC	CALCULATED
CMF	2 digit indicator of Career Management Field	CALCULATED
*Destination	Calculated from MPAORG for loss data (OUT, Retired, Military, TPU)	CALCULATED

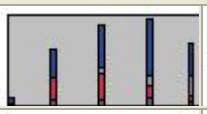
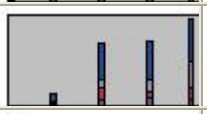
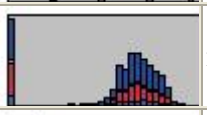
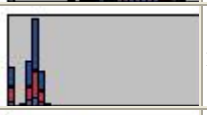
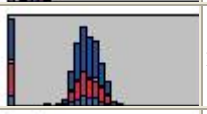
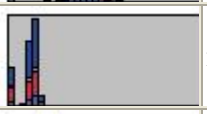
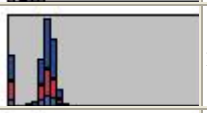
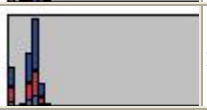
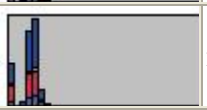
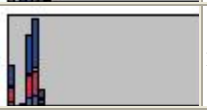
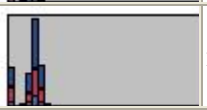
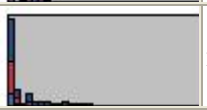
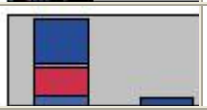
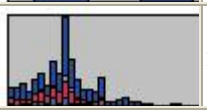
APPENDIX B DATA AUDIT

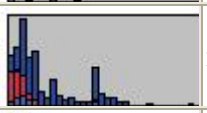
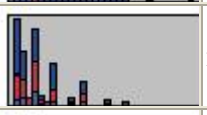
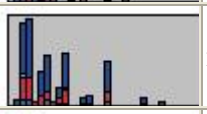
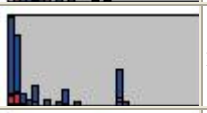
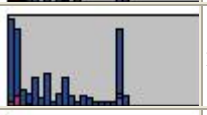
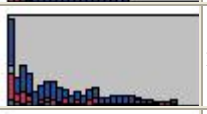
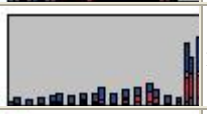
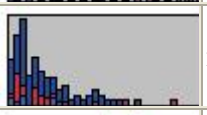
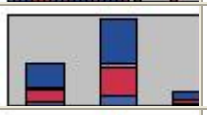
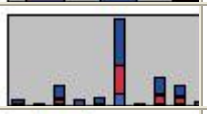
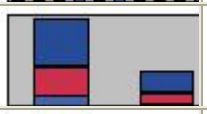
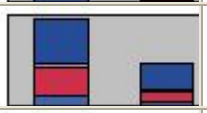
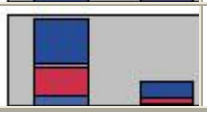
Table 11. Data Audit of 82 input fields

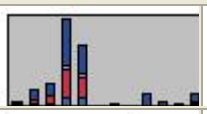
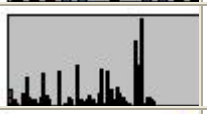
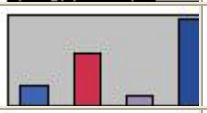
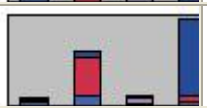
	Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique
1	RSN_FLAG		set	--	--	--	--	--	14
2	DtBiY		range	1175	1989	1974.539	9.774	-2.518	
3	HIGH		range	0.000	99.000	67.231	8.899	-6.235	
4	WGNT		range	0.000	999.000	165.750	42.310	-0.835	
5	HT_WT_IND		set	--	--	--	--	--	4
6	SEX		set	--	--	--	--	--	4
7	ETH_GP		set	--	--	--	--	--	25
8	RACE		set	--	--	--	--	--	7
9	CITZ		set	--	--	--	--	--	9
10	MAST		set	--	--	--	--	--	10
11	DEPN		range	0	40	0.988	1.352	1.476	
12	RELI		set	--	--	--	--	--	163
13	CIVO		range	0	999	533.594	397.846	0.049	

14	STATE		set	--	--	--	--	--	110
15	GRAS		set	--	--	--	--	--	22
16	GRADE		set	--	--	--	--	--	22
17	DtPEBD		range	0	2006	1995.443	9.518	-53.629	
18	DtEXP		range	1900	2026	2002.742	9.479	-1.240	
19	DtTPUEXP		range	1901	2060	2007.346	4.895	1.647	
20	DtLRA		range	1920	2028	1997.207	7.470	-1.447	
21	ACT_FEDSVC		range	0.000	3173.00	172.226	293.046	2.734	
22	PPSC		range	111111	444444	115714.5	27519.048	6.699	
23	PHCC		set	--	--	--	--	--	19
24	APFT_IND		set	--	--	--	--	--	4
25	DEPL		set	--	--	--	--	--	48
26	MILED_COMP		set	--	--	--	--	--	37
27	CIED		set	--	--	--	--	--	29
28	CVEL		set	--	--	--	--	--	28
29	DtENTRY		range	0	2006	1994.646	22.522	-74.386	

30	DtENTRES		range	0	2006	1971.179	223.684	-8.687	
31	AFSG		set	--	--	--	--	--	15
32	AFQT		range	0.000	99.000	55.636	24.233	-0.340	
33	DtETS		range	0	2080	2007.965	23.870	-80.367	
34	NEXE		range	1	9	1.361	0.803	3.529	
35	CMXT		range	1	48	15.183	10.874	1.265	
36	SKLVL		set	--	--	--	--	--	149
37	GOOD_YRSVC		range	0.000	40.000	5.353	7.241	1.641	
38	UCAG		set	--	--	--	--	--	72
39	FCPSCD		set	--	--	--	--	--	5
40	SOPTDD		set	--	--	--	--	--	3
41	MILSPI		set	--	--	--	--	--	3
42	DtADTE		range	199	2006	1997.714	6.224	-70.323	
43	DMOSQ		set	--	--	--	--	--	9
44	TIER		set	--	--	--	--	--	13
45	PRIX		range	0	5	1.627	1.269	0.978	

46	SECX		range	0	5	2.845	1.329	0.086	
47	ALTX		range	0	5	3.575	1.191	-0.502	
48	ASVABCL		range	0.000	191.000	86.706	42.597	-1.361	
49	ASVABCO		range	0.000	920.000	85.094	42.328	-1.239	
50	ASVABEL		range	0.000	311.000	85.414	42.270	-1.309	
51	ASVABFA		range	0.000	999.000	86.714	42.976	-1.231	
52	ASVABOF		range	0.000	611.000	85.161	42.111	-1.321	
53	ASVABGT		range	0.000	910.000	86.517	41.831	-1.330	
54	ASVABGM		range	0.000	945.000	84.317	42.180	-1.205	
55	ASVABMM		range	0.000	980.000	84.023	41.937	-1.215	
56	ASVABSC		range	0.000	980.000	84.867	42.057	-1.277	
57	ASVABST		range	0.000	910.000	86.380	42.655	-1.293	
58	FirstOfBonus		range	0	60000	2047.587	3632.123	2.232	
59	DEPPER		flag	--	--	--	--	--	2
60	SumOfDURATION		range	1.000	1812.000	431.350	228.957	0.861	
61	LastOfAPC_DESC		set	--	--	--	--	--	7

62	CombatFLG		flag	0.000	1.000	--	--	--	2
63	CNTDPLY		range	1	11	1.237	0.551	3.102	
64	Total Of Amount		range	7.000	40000.0	5432.237	5163.847	1.404	
65	AMT02		range	40.000	8000.00	1052.235	766.140	1.425	
66	AMT03		range	3.000	6000.00	1037.393	739.661	1.615	
67	AMT04		range	58.000	9400.00	1066.047	796.501	2.072	
68	AMT05		range	23.000	31600.0	3476.110	4937.610	1.762	
69	AMT06		range	69.000	30900.0	4980.207	5382.329	1.037	
70	BasMGIB		range	0.000	14406	2953.404	2609.825	0.776	
71	DtMGIB		range	1985	2007	2001.238	5.517	-1.040	
72	KicMGIB		range	0.830	14513	2018.870	1974.481	1.954	
73	Partition		set	--	--	--	--	--	3
74	CIVO_BIN		set	--	--	--	--	--	10
75	Profile		flag	--	--	--	--	--	2
76	Mobilized		flag	--	--	--	--	--	2
77	Deployed		flag	--	--	--	--	--	2

78	RRC		set	--	--	--	--	--	12
79	CMF		set	--	--	--	--	--	76
80	Destination		set	--	--	--	--	--	4
81	$C-Destination		set	--	--	--	--	--	4

LIST OF REFERENCES

Hansen, Michael L. and MacLeod, Ian D., *Retention in the Reserve and Guard Components*, Alexandria, VA, CNA, April 2004.

Hosek, James R. and Totten, Mark, *Serving Away From Home: How Deployments Influence Reenlistment,* Santa Monica, CA, RAND, 2002.

Dolfini-Reed, Michelle A. and Parcell, Ann D., *Determining Patterns of Reserve Attrition Since September 11, 2001,* Alexandria, VA, CNA, June 2005.

Hand, David and Mannila, Heikki and Smyth, *Principles of Data Mining,* Cambridge, MA, MIT Press, 2001.

SPSS Inc., *Clementine 10.0 User's Guide,* Chicago, IL, SPSS Inc., 2005.

SPSS Inc., *Clementine 10.0 Node Reference,* Chicago, IL, SPSS Inc., 2005.

USARC, *Data Reference Guide,* FT MacPherson, GA, USARC, 2006.

USAREC, *Personal Communications,* FT Knox, KY, 2006.

RTD, *Personal Communications,* Atlanta, GA, 2006.

Ginther, T.A., *Army Reserve Enlisted Aggregate Flow Model,* M.S. Thesis, Naval Postgraduate School, California, June 2005.

Brau, J.W., *Improving the Quality And Personnel Fill Rates of U.S. Army Reserve Units,* M.S. Thesis, Naval Postgraduate School, California, June 2004.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, VA

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, CA

3. Department of the Army
   Office of the Chief of the Army Reserve
   Washington, DC

4. Dr. Samuel E. Buttrey
   Naval Postgraduate School
   Monterey, CA

5. Dr. Roberto Szechtman
   Naval Postgraduate School
   Monterey, CA

6. MAJ Alison Godfrey
   OCAR - PAE
   Washington, DC

7. MAJ Alfred Evans
   RTD
   Atlanta, GA